Original Paper

Fairness in Mobile Phone–Based Mental Health Assessment Algorithms: Exploratory Study

Jinkyung Park[1], MA; Ramanathan Arunachalam[2], MSc; Vincent Silenzio[3], MD; Vivek K Singh[1,4], PhD

[1]School of Communication & Information, Rutgers University, New Brunswick, NJ, United States

[2]Department of Computer Science, Rutgers University, New Brunswick, NJ, United States

[3]School of Public Health, Rutgers University, Newark, NJ, United States

[4]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, United States

Corresponding Author:
Vivek K Singh, PhD
School of Communication & Information
Rutgers University
4 Huntington Street
New Brunswick, NJ, 08901
United States
Phone: 1 848 932 7588
Email: v.singh@rutgers.edu

Abstract

Background: Approximately 1 in 5 American adults experience mental illness every year. Thus, mobile phone–based mental health prediction apps that use phone data and artificial intelligence techniques for mental health assessment have become increasingly important and are being rapidly developed. At the same time, multiple artificial intelligence–related technologies (eg, face recognition and search results) have recently been reported to be biased regarding age, gender, and race. This study moves this discussion to a new domain: phone-based mental health assessment algorithms. It is important to ensure that such algorithms do not contribute to gender disparities through biased predictions across gender groups.

Objective: This research aimed to analyze the susceptibility of multiple commonly used machine learning approaches for gender bias in mobile mental health assessment and explore the use of an algorithmic disparate impact remover (DIR) approach to reduce bias levels while maintaining high accuracy.

Methods: First, we performed preprocessing and model training using the data set (N=55) obtained from a previous study. Accuracy levels and differences in accuracy across genders were computed using 5 different machine learning models. We selected the random forest model, which yielded the highest accuracy, for a more detailed audit and computed multiple metrics that are commonly used for fairness in the machine learning literature. Finally, we applied the DIR approach to reduce bias in the mental health assessment algorithm.

Results: The highest observed accuracy for the mental health assessment was 78.57%. Although this accuracy level raises optimism, the audit based on gender revealed that the performance of the algorithm was statistically significantly different between the male and female groups (eg, difference in accuracy across genders was 15.85%; *P*<.001). Similar trends were obtained for other fairness metrics. This disparity in performance was found to reduce significantly after the application of the DIR approach by adapting the data used for modeling (eg, the difference in accuracy across genders was 1.66%, and the reduction is statistically significant with *P*<.001).

Conclusions: This study grounds the need for algorithmic auditing in phone-based mental health assessment algorithms and the use of gender as a protected attribute to study fairness in such settings. Such audits and remedial steps are the building blocks for the widespread adoption of fair and accurate mental health assessment algorithms in the future.

(JMIR Form Res 2022;6(6):e34366) doi: 10.2196/34366

KEYWORDS

algorithmic bias; mental health; health equity; medical informatics; health information systems; gender bias; mobile phone

Introduction

Background

Various machine learning (ML) algorithms are increasingly being used to make crucial decisions previously made by humans. Whether they are involved in approving loans, granting college admissions, or identifying the need for additional health support, automated algorithms find patterns, predict outcomes, and make decisions that may have consequential impacts on individuals' lives [1]. Indeed, the dependency on algorithms has eased our lives by replacing subjective human decisions with ML algorithms. The movement toward the application of automated algorithms in the health domain was not an exception. For instance, the proactive assessment of an individual's mental health is essential for maintaining a healthy and well-functioning society [2]. Although this holds the promise of dramatically wider access to mental health care, it is also fraught with inequities that can often inadvertently be baked into the algorithmic prediction of mental health levels.

ML algorithms attempt to find the generalized pattern from the training data, and sometimes these algorithms can manifest inherent biases across demographic characteristics such as age, race, ethnicity, and gender. A reason for the existing biases can be explained by *negative legacy* [3] (ie, the absence of sufficient data for a particular demographic group). For example, giving loans mostly to higher-income groups in the past may result in disapproval of loans to lower-income groups by algorithms that were informed by historical data, resulting in potential damage to individuals belonging to lower-income groups.

Such biases can be especially deleterious if they are part of health care algorithms. For instance, a recent study by Allen et al [4] found that algorithms used to assess mortality scores exhibit differential accuracy across races, thereby increasing racial disparities in health care. Similarly, Gianfrancesco et al [5] demonstrated that algorithmic predictions based on electronic health records can discriminate against multiple demographic groups. In particular, Obermeyer et al [1] showed that existing algorithms do not adequately identify the need for health support for people of color.

Building on these trends, we move the discussion of algorithmic fairness to mobile mental health assessment algorithms, which have been increasingly used in recent times [6]. With >6 billion users, mobile phones are one of the most ubiquitous consumer devices in the world. Many of them (especially smartphones) have capabilities conducive to monitoring an individual's physical activity, location, and communication patterns, each of which has been connected to mental health in the past [7,8]. Thus, mobile phones hold significant promise as a platform for monitoring multiple indicators of mental health risks and improving long-term management and treatment delivery to people with mental health issues [7,9]. At the same time, the creation of phone data–based ML models without considering the aspects of justice and fairness could reify, amplify, and multiply existing health disparities for certain segments of society (eg, women). Considering the abovementioned factors, the main research questions (RQs) of this study were as follows:

- RQ1: Are mobile phone-based mental health algorithms susceptible to bias in terms of gender?
- RQ2: Is it possible to reduce the level of bias while maintaining high accuracy?

Related Work

Predicting Mental Health

Over the past few decades, mental health has typically been assessed based on self-reported surveys that involved sporadic sampling, most of which were initiated after some significant events had taken place in an individual's life. Recently, as the availability of mobile phone data has increased, several studies have suggested using mobile phone data to detect and predict mental health conditions. Wang et al [10] introduced a mobile phone sensing system to automatically infer mental well-being, including depression, stress, flourishing, and loneliness. The study reported that automatically sensed conversation, activity, mobility, and sleep were significantly associated with mental health outcomes. By collecting data from sensors in mobile phone users (eg, location, messages, and calls), a longitudinal study showed a relationship between users' routines and moods [11]. Another study also found that mobile phone–based features such as call count, call response rate, and the number of new contacts are positively associated with mental health [8]. Using location information collected by a mobile phone app, Canzian and Musolesi [12] assessed the correlation between mobility patterns and the presence of depressive mood. A similar study also presented the relationship between depressive symptoms and the use of mobile phones and the movement through geographic spaces [7].

The results of the abovementioned studies provide clear evidence of interconnections between mobile phone data features and mental health conditions. More importantly, they suggested the potential of developing phone-based ML algorithms as a basis for the unobtrusive prediction of mental health conditions. However, to the best of our knowledge, no study has examined the possibility of algorithmic bias in predicting mental health status by using mobile phone data. Motivated by previous work on algorithmic fairness (see the *Algorithmic Fairness* section), this study attempted to mitigate the discriminatory impact of gender on mental health prediction algorithms.

Algorithmic Fairness

An increasing amount of research has suggested that ML algorithms in many domains are not free from discriminatory decision-making. Even with the best intentions, data-driven algorithmic decision-making processes can reproduce, inherit, or reflect the existing social biases. Algorithmic bias may stem from different sources, including (1) input data that may have unequal representation from different groups, (2) an algorithm that has been inadvertently or knowingly coded to make unfair decisions, (3) misuse of certain models in a different context, and (4) biased training data, which reaffirms that social biases may be used as evidence that an algorithm performs well [13]. Broadly, the sociotechnical system framework underscores that the value system of the algorithm developers is coded during the algorithm design process; hence, each assumption (often implicit) made by the developers influences the real-world performance of the algorithm [14].

At the same time, multiple bias mitigation techniques have been developed for fairness in the ML literature [15,16]. Roughly, they attempt to counter such algorithmic bias by modifying the training data (preprocessing), learning algorithms (in-processing), or prediction (postprocessing). Preprocessing approaches focus on adapting the data going into the algorithms [16], in-processing approaches change the core algorithm (eg, change optimization function) [15], and postprocessing algorithms tend to modify the predicted labels to increase fairness [17].

Despite the plethora of related work, attempts to ensure algorithmic fairness toward a protected attribute (gender in our case) in the algorithmic assessment of mental health (high or low) have not been made.

Gender Bias

Various attempts have been made to tackle the issue of gender bias in computer algorithms by auditing algorithms for gender bias and modifying algorithms to eliminate stereotypes. For example, a study found that image search results for occupations could amplify gender stereotypes by portraying the minority gender as less professional [18]. Another study found gender stereotypes in word embeddings (eg, a framework to represent text data as vectors) and created debiasing algorithms to reduce gender bias while preserving the utility of the embeddings [19]. Furthermore, Zhao et al [20] tackled the problem of the effect of data imbalance, arguing that such data imbalance can worsen discrimination in terms of gender. They quantified the biases in visual recognition models and calibrated the models to reduce bias. However, no research has been conducted on gender equality using classification algorithms that predict mental health.

This study addressed the problem of identifying and reducing gender bias, as the overrepresentation of men in training data could accelerate gender inequality in mental health prediction algorithms. Particularly, we focused on the issue of *negativelegacy*, as suggested by Kamishima et al [3], which involves the idea that unfair sampling or labeling in the training data may lead to a disparate impact [16,21] on a certain group of people (eg, granting loans mostly to those who with higher income in the past may result in disapproval of loans to those with low income by the algorithms).

Perspective on Fairness and Justice

There exist multiple interpretations of fairness in the algorithmic fairness literature [22]. For instance, scholars define fairness in terms of maximizing utility for groups or respecting various rules such as individual rights and freedoms [23,24]. However, other interpretations abound, some of which are mutually incompatible [25].

The most commonly used approaches are those based on *distributive* and *procedural* justice [22]. While distributive justice focuses on how outcomes are distributed across the population, procedural justice focuses on the processes used to undertake the decisions [26,27].

An influential philosophical theory of fairness is attributed to the 20th-century philosopher Rawls, who equated fairness and justice, arguing broadly that fairness is *a demand for impartiality* [21,22]. In this study, we followed the approach for distributive justice based on the interpretation of Rawls. Specifically, we considered an algorithm to be fair if its performance did not vary for individuals with different demographic descriptors (eg, gender).

This is related to the concept of *disparate impact* [28]. Disparate impact, in US labor law, refers to practices in areas such as employment and housing, which affect one group of people of a protected characteristic more adversely than another, even when the rules applied by employers or landlords appear to be neutral [29]. Most federal civil rights laws protect against disparate impacts based on race, color, religion, national origin, and sex as protected traits, and some laws include disability status and other traits.

Methods

Data Set

We used a labeled data set from a previous study by Singh and Long [8], which explored the associations between call log data and mental health based on a 10-week field and laboratory study. The data set included phone-based behavioral data and self-reported mental health survey data. Phone-based data (eg, call volume, interaction dynamics, diversity in contacts, tie strength, and temporal rhythms) were collected through the app installed on each participant's mobile phone. Meanwhile, mental health was measured via in-person survey sessions using the Mental Health Inventory subscale of the 36-Item Short Form Health Survey [30]. After passing a preprocessing and classification process, the study showed that automated ML algorithms using phone-based features achieved up to 80% accuracy in automatically classifying the mental health level (above or below the mean) of an individual [8].

A total of 59 participants completed the survey administered by Singh and Long [8]. However, some participants did not complete all the surveys, and some did not enter the correct identifier (International Mobile Equipment Identity [IMEI] number) consistently across surveys. This resulted in a subset of 45 participants in the study [8]. For this study, we returned to the survey data and decided to manually handle the *off-by-one* errors (ie, the mismatch in IMEI for different surveys only by 1 digit). Given that IMEI numbers have 14 to 15 digits, in the approximately 60-participant sample size, we considered the odds of 2 participants to be off by just 1 digit without human error being extremely low. This process helped us obtain a complete data set (ie, phone data, a mental health survey, and a demographic survey) for 55 participants.

The data set we obtained from Singh and Long [8] had gender as a demographic attribute that we considered a protected attribute. Note that a protected attribute in the algorithmic fairness literature is one on which performance should not depend [15]. Among these 55 participants, 21 (38%) self-reported their gender as women or female (minority class), and 34 (62%) self-described as men or male. Note that this study does not differentiate between (biological) sex and (socially construed) gender. In addition, note that we consider the use of binary gender as a limitation of this study. Future studies should be conducted, which include participants with nonbinary gender identities.

Preprocessing and Model Training

The initial obtained data set was imbalanced (ie, there was not enough data for one class), which is a common problem in the fairness literature [31]. To mitigate the effect of imbalance, we applied the synthetic minority oversampling technique [32] to the training data (the test data remained in the original ratio). This technique works in balancing the data set by generating synthetic observations based on the existing minority observations.

Before moving on to the application of any ML algorithm, the missing values were filled with the median values of the corresponding features. To reduce the impact of features with high variance, the features were standardized by removing the mean and scaling to unit variance. To build a classification model for high or low mental health scores, instances were labeled into 2 categories (1=high and 0=low) via a median split.

With small sample data and high-dimensional space, there is always a chance of overfitting and reduced generalization. To avoid these issues, we used principal component analysis [33]. Principal component analysis confirmed that the top 5 components explained >99% of the variance (the larger the variation across a dimension, the more the information it contains); hence, we used the top 5 components as features for model creation.

The abovementioned latent features were passed to several classification algorithms to classify the level of mental health (ie, whether the score was above or below the mean score of the population). As the sample data size was relatively modest, we refrained from splitting the data set into training and test sets. Instead, as suggested by prior literature [8,34], we applied 5-fold cross-validations and experimented with 5 popular classification algorithms, including logistic regression, support vector machine, random forest, k-nearest neighbors, and multilayer perceptron neural networks using the *scikit-learn* library [35]. We ran all algorithms for 100 iterations, and the results are reported in the form of average overall accuracy, male accuracy (ie, accuracy for male individuals), and female accuracy (see the *Results* section).

Using the abovementioned data, we could, in principle, replicate the approach described by Singh and Long [8]. Although the features used were the same, we must note that the implementation was undertaken de novo with different preprocessing steps.

Auditing Mental Health Algorithms for Bias

Gender was selected as a protected attribute. Following the previous literature [36,37], men were considered the privileged group, and women were considered the unprivileged group. As there are multiple metrics to characterize accuracy in traditional ML (eg, observed accuracy, precision, recall, and F_1 score), past literature has discussed the need for multiple metrics to characterize bias in ML [13,31]. In this study, we adopted the five most commonly used metrics [15,16,38]:

1. Delta accuracy captures the difference in the accuracy of samples belonging to privileged and unprivileged groups based on sensitive features (eg, gender and race or ethnicity).
2. Delta true positive rate (ΔTPR) focuses on equal opportunity for truly deserving entries in both privileged and unprivileged groups to obtain a positive label (eg, higher mental health label) from the algorithm [13,15].
3. Delta false positive rate (ΔFPR) ensures that both the true positive rate and the false positive rate (instances where undeserving candidates are granted positive outcomes) are equal across different groups [15,39].
4. Statistical parity difference (SPD) calculates the difference in the probability of favorable outcomes from the algorithm being obtained by the unprivileged group to that of the privileged group [38].
5. Disparate impact captures the ratio of the probability of favorable outcomes for the unprivileged group to that of the privileged group [16] (see Multimedia Appendix 1 [13,15,16,39-41] for more details on the 5 metrics).

Following the principle of disparate impact, a fair information system is one in which the performance does not vary for individuals with different demographic descriptors (eg, gender); hence, the disparate impact metric should be close to 1.0. However, for practical settings, a model is considered biased if its value is <0.8 [40]. Meanwhile, the values of delta accuracy, ΔTPR, ΔFPR, and SPD should be close to zero in fair systems. Following the previous literature [39,41], we used a 2-tailed *t* test to assess whether there was a significant difference in accuracy, true positive rate, and false positive rate levels observed for the privileged and unprivileged groups.

Reducing Algorithmic Bias in Mental Health Assessment

Disparate impact remover (DIR) [16] is a preprocessing algorithm that modifies the feature values of the data set and makes the algorithm discrimination aware at the time of training. It does not require any changes in the classification algorithm, nor does it amend or postprocess the results of the classification algorithm, thus making it robust and applicable to different algorithms. The scenario in which DIR is needed to preprocess the data set depends on the metric called *balanced error rate (BER)*, defined as follows:

BER = (error rate [S = privileged] – error rate [S = unprivileged]) / 2

In algorithmic fairness, the notion of BER is more important than the notion of traditional accuracy as, in most data sets, the contribution of the underprivileged attribute to the entire data set is lesser than that of the privileged attribute. For example, let us consider a data set with 100 rows, where 90 rows belong to the privileged group and 10 rows belong to the unprivileged group. With this data set, if the algorithm predicts all privileged rows right and unprivileged wrong, the error rate would be 10/100, which is 0.1, whereas the BER would be (0+1)/2, which is 0.5.

An approach discussed in the literature [16,17] is to replace the raw values of the data features with normalized variants that capture how extreme the value for an individual (eg, female) stands out within their own demographic group (eg, other women). In particular, the approach suggested by Feldman et al [16] tackles this issue by allowing the considered classes to have equal probabilities of scoring high values for any of the chosen features. With a toy example, where output is college admissions, input is Scholastic Assessment Test (SAT) scores, and with a binary notion of gender (men and women) for the protected class, this approach gives men and women separate scores based on their ranking within their own genders. For example, a man with an 80th percentile SAT score within the men's group is considered just as worthy as a woman with an 80th percentile SAT score within the women's group, irrespective of the actual SAT scores. In this way, the approach supports an equitable admission process across 2 genders. Note that in many practical settings, it is useful to undertake *partial repairs* (eg, move the scores at the same percentile across the privileged and unprivileged groups to be closer to each other rather than being congruent). Finally, the above approach can be extended to multidimensional input features for the algorithm. In the considered domain (phone-based mental health assessments), phone use patterns for men and women are known to differ [42,43]. Hence, using the same thresholds for the features (eg, number of phone calls) of men and women as symptoms of mental health issues could yield erroneous and biased results.

In this study, the DIR algorithm for bias reduction was implemented in Python using the IBM AIF360 library [15]. The algorithm was run 100 times, with each iteration having a shuffled version of the data set. The average results for the accuracy and fairness metrics are presented in the *Results* section.

Results

Mental Health Assessment Results

Table 1 shows the accuracy of multiple well-known ML algorithms for men and women (averaged over 100 iterations). The best-performing algorithm was random forest, which yielded 78.57% accuracy. These results are similar but not the same as those described by Singh and Long [8]. In both studies, the random forest algorithm yielded the best performance, and the highest observed accuracy was close to 80%. The random forest model with the highest accuracy had 100 estimators or number of trees in the forest and a maximum depth of 6.

Table 1. Results showing the average overall accuracy, accuracy for men, and accuracy for women for various machine learning models in mental health assessment (averaged over 100 iterations).

Machine learning models	Overall accuracy (%), mean (SD)	Male accuracy (%), mean (SD)	Female accuracy (%), mean (SD)	Delta across gender (%), mean (SD)	*P* value of the 2-tailed *t* test on delta
Multilayer perceptron neural networks	59.99 (3.67)	58.68 (8.14)	61.92 (9.24)	12.10 (10.41)	<.001
Support vector machine	63.17 (2.91)	65.98 (6.49)	59.60 (8.37)	12.20 (8.67)	<.001
Logistic regression	58.48 (2.69)	66.59 (5.47)	47.38 (6.75)	19.73 (9.80)	<.001
K-nearest neighbors	61.77 (1.78)	70.43 (3.72)	49.63 (5.89)	20.96 (8.46)	<.001
Random forest	78.57 (1.61)	87.16 (2.73)	71.31 (2.51)	15.85 (0.22)	<.001

Audit Results

We compared the accuracies of different algorithms for the male and female groups (Table 1). The performance was found to be significantly different for the 2 groups in each of the considered algorithms based on a nonpairwise (2-tailed) *t* test (α=.05; *P*<.001) [41]. This indicates that the commonly used ML algorithms, when used for phone-based mental health assessment, are susceptible to bias.

There, a trade-off is expected between accuracy and fairness (ie, with increased fairness, there is typically a dip in accuracy) [31], the random forest model with the highest observed accuracy was selected as the baseline model for further inspection of fairness.

For random forest, the average absolute delta accuracy was 15.85% (Table 2). The absolute values of ΔTPR and ΔFPR were 0.88% and 33.43%, respectively. The average SPD was 26.1%, and the average disparate impact was 0.682, which were distant from the ideal values of 0 and 1.0, respectively.

Table 2. The average score for bias metrics in the random forest–based mental health assessment algorithm (average of 100 iterations).

Bias metrics	Observed score, mean (SD)	Ideal score
Delta accuracy (%)	15.85 (0.22)	0
Delta true positive rate (%)	−0.88 (8.39)	0
Delta false positive rate (%)	33.43 (13.50)	0
Statistical parity difference (%)	26.1 (4.16)	0
Disparate impact	0.682 (0.049)	1.0

For 4 of the 5 considered metrics (ie, except ΔTPR), the fairness scores were far from the ideal scores. In other words, the developed model yielded significantly different outcomes for individuals across genders despite reasonable aggregate performance. More precisely, the model was mostly biased against the unprivileged group (in this case, women), and the disparate impact appeared to be a major issue.

Bias Reduction Results

We recomputed the abovementioned bias metrics after applying the bias reduction algorithm (DIR), and the results averaged over 100 iterations are reported in Table 3. Furthermore, a comparison of the results before and after applying the bias reduction algorithm is presented in Table 4.

Table 3. The average score for bias metrics after applying the disparate impact remover approach (average of 100 iterations).

Bias metrics	Observed score, mean (SD)	Ideal score
Delta accuracy (%)	1.66 (1.56)	0
Delta true positive rate (%)	3.74 (6.74)	0
Delta false positive rate (%)	5.58 (9.88)	0
Statistical parity difference (%)	−2.70 (1.71)	0
Disparate impact	1.09 (0.041)	1.0

Table 4. Comparison of delta accuracy, statistical parity difference, and disparate impact before and after applying the postprocessing algorithm.

Bias metrics	Baseline model, mean (SD)	After bias reduction, mean (SD)	Difference	*P* values of 2-tailed *t* test on delta
Delta accuracy (%)	15.85 (0.22)	1.66 (1.56)	14.19	<.001
Delta true positive rate (%)	−0.88 (8.39)	3.74 (6.74)	4.63	<.001
Delta false positive rate (%)	33.43 (13.50)	5.58 (9.88)	27.85	<.001
Statistical parity difference (%)	26.10 (4.16)	−2.70 (1.71)	28.80	<.001
Disparate impact	0.682 (0.049)	1.09 (0.041)	0.408	<.001

To test the significance of these improvements, we conducted a 2-tailed *t* test with α=.05 for each of the bias metrics for the before and after scores. The changes in all metrics were noteworthy (*P*<.001). The bias levels were reduced for 4 of the 5 metrics considered in this study. The only exception was ΔTPR, which was the only metric with a low (<5%) score in the baseline condition. This value remained <5% before and after the bias reduction process.

Note that as we move toward making the algorithm less biased, there is often a trade-off that arises in the form of the reduced overall accuracy of the model [13]. The accuracy levels for men and women were 87.16% and 71.31%, respectively (Δaccuracy 15.85%; mean 78.50%), before bias reduction. The accuracy levels changed to 78.49% and 76.83% for men and women, respectively (Δaccuracy 1.66%; mean 76.83%), after the bias reduction process. The 1.38% reduction (78.50%-77.12%) in the model accuracy was considered an acceptable loss in accuracy for the abovementioned improvements in fairness.

Discussion

Principal Findings

RQs of the Study

The first RQ in this work was as follows: are mobile phone–based mental health algorithms susceptible to bias in terms of gender?

As summarized in Table 1, we found statistically significant differences across genders in the performance of phone-based mental health assessment algorithms with an array of common ML algorithms. All of these point to the potential for disparate impact across gender with mental health assessment algorithms.

With respect to the performance of the highest accuracy algorithm (using random forest), we found noticeable differences in the performance of the algorithm across genders via the 5 commonly used bias metrics. As shown in Table 2, there was a difference in terms of all 5 metrics between the male and female groups. In particular, we found that the disparate impact ratio was 0.682 in the initial model. However, this value was much lower than the often recommended (and legally accepted) threshold of 0.8, irrespective of the intent of the designers [29]. Although the in-principle replications of algorithms described in the past literature may yield reasonable accuracy, their deployment will require them to meet the legal and ethical guidelines of disparate impact. In addition, similar fairness issues have been well studied in some other spaces (eg, policing and bank loans [44,45]); they are much less explored in algorithmic mental health assessment. However, they will become important with the increased deployment or adoption of mobile mental health tools.

The results also point to another domain in which women are disadvantaged. As per the US Department of Labor Statistics, women earn 82 cents for every dollar earned by men [46]. Similarly, recent research has reported worse performance for women in face recognition [47], Google Translate [48], and image search results [18]. The awareness of such disparities is an important first step in the creation of countermeasures. Broadly, such results in intersection with growing movements such as *Data Feminism* [49] can support the creation of more equitable algorithms. Specifically, we hope that our findings will shed light on the need to ensure fairness in emerging mental health–related domains.

Finally, there are multiple potential reasons for the reduced performance of women in the considered algorithms. Given that the performance is consistently poorer for all the considered ML algorithms (Table 1), possible explanations may lie in the *negative legacy* and *data set imbalance*. Data imbalance is the lack of data samples from a particular demographic group for algorithms to learn from, and negative legacy refers to the lack of positive examples for algorithms to learn from for the unprivileged group [13,31]. For instance, Buolamwini and Gebru [47] argued that a lack of training samples is a reason for poorer performance for women and people of color. Similar to other areas, and perhaps even more urgently, there is a need for more diverse data samples to create accurate and fair ML models in mental health assessment algorithms.

The second RQ in this study was as follows: is it possible to reduce the level of bias while maintaining high accuracy?

On the basis of the results summarized in Table 4, we found that the DIR approach was effective in reducing the disparity in the performance of phone-based mental health assessment algorithms across genders. As reported in Table 4, there were statistically significant differences in terms of all 5 fairness metrics considered upon the application of the DIR approach.

Past literature has discussed the need for multiple metrics to characterize bias in ML [13,31] and that metrics can be orthogonal to each other [25,44]. A suggested process is for system designers to identify a set of parameters that they consider appropriate for a given task [50]. In this study, we considered disparate impact to be an important criterion, considered in consultation with the scores for other fairness metrics. In the considered scenario, noticeably large reductions in bias levels were observed regarding the 4 metrics, except for ΔTPR, where the scores were <5% before and after bias reduction. Finally, we noted that there was a 1.38% decrease in accuracy upon the application of the bias reduction approach.

Overall, we interpreted the results to imply that it is often possible to create fairer versions of algorithms. However, given the variety of fairness metrics that can be considered and the complexities of practical scenarios, the process of bias reduction is likely to involve a human-in-the-loop process and consideration of the trade-offs in terms of multiple metrics [50]. Hence, rather than identifying a silver bullet solution, there might be opportunities for multiple small modifications that allow fairer versions of the algorithms. Having said that, value-sensitive design needs to be an important part of the future design of similar applications [51], and algorithmic audits need to become an essential step in the process of medical approval of newer (algorithmic) diagnostic tools.

The obtained results have multiple implications for different stakeholders engaged in health information systems.

Health Informatics Researchers and Policy Designers

This study moves the conversation with health policy designers beyond the equity of the built environment (eg, access to hospitals and parks) to the equity of data infrastructure, which can profoundly influence the health outcomes for millions of individuals going forward [52]. Although there exist multiple legal and policy guidelines that counter the physical aspects of bias (eg, redlining [53]), there is relatively little work on legal and policy frameworks with digital algorithms that undertake similar roles.

Health Care Technology Companies

This study identified a feasible pathway for creating algorithms that balance accuracy and equity in the creation of novel health care applications. Hence, the findings support the creation of equitable versions of just-in-time mobile mental health intervention apps.

Health Care Providers

This study allows for more robust detection and flagging of mental health issues in patients. Fairer algorithms will reduce the odds of patients being flagged for interventions incorrectly simply because of demographic characteristics, thus allowing for the better alignment of resources between individual providers and the health care industry at large.

The Public

The ultimate goal of this study was to create and promote equity in mental health information technology. The fairness of algorithms is intimately connected with trust and adoption. In fact, recent research suggests that disparate impact diminishes consumer trust, even for advantaged users [40]. A robust fair detection process will allow for the scalable delivery of just-in-time and tailored mental health support services to a wider population. This is important, given the huge disparity between the need for mental health support and the percentage of the population that uses mental health services [54].

Limitations

This study has some limitations. It focused on a single data set with 55 individuals and considered a specific type of feature (phone data based, as described by Singh and Long [8] in the past literature). The use of binary gender in the assessment is another limitation of this study. Although this study examined many of the commonly used ML methods, other approaches are well represented in the literature. Hence, we will be cautious in generalizing the results until they are supported at a scale with samples of more representative populations and many other ML algorithms. Future work may also suggest other bias reduction techniques to reduce the discriminatory outcomes of mental health assessment algorithms based on protected attributes. At the same time, this work is the first empirical effort to analyze the difference in the performance of mental health assessment algorithms based on gender. A key contribution of this study is the motivation for future work in this domain using varied data sets and methods.

Conclusions

This study grounds the use of gender as a protected attribute to study fairness in phone-based mental health assessment algorithms. Mobile phones are now actively used by billions of individuals; hence, the automatic assessment of mental health using ML algorithms could potentially be beneficial in estimating and intervening in billions of individuals' mental health conditions. An audit of commonly used ML algorithms for mental health assessment revealed that the performance of these algorithms can vary significantly depending on gender. This disparity in performance was found to be noticeably reduced after the application of a DIR approach by adapting the data used for modeling. The results move the literature forward on fairness in mental health assessment algorithms, particularly with gender as a protected attribute. Future work could consider larger data sets, protected attributes other than gender, and a newer approach to creating fair and accurate mental health assessment algorithms. Such results will pave the way for accurate and fair mental health support for all sections of society.

Acknowledgments

The authors would like to thank Gautam Sikka and Rahul Ellezhuthil for their help with the data analysis. This work was partly based on funding from the Rutgers Community Design for Health and Wellness Interdisciplinary Research Group, Rutgers Center for COVID-19 Response and Pandemic Preparedness, and Rutgers School of Communication & Information Scholarly Futures Grant.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The 5 metrics to measure bias in machine learning algorithms.
[DOCX File , 21 KB-Multimedia Appendix 1]

References

1. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
2. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Ment Health 2020 Nov;23(4):161-166. [doi: 10.1136/ebmental-2020-300180] [Medline: 32998937]
3. Kamishima T, Akaho S, Asoh H, Sakuma J. Fairness-aware classifier with prejudice remover regularizer. In: Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases. 2012 Presented at: ECML PKDD '12; September 24-28, 2012; Bristol, UK p. 35-50 URL: https://doi.org/10.1007/978-3-642-33486-3_3 [doi: 10.1007/978-3-642-33486-3_3]
4. Allen A, Mataraso S, Siefkas A, Burdick H, Braden G, Dellinger RP, et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. JMIR Public Health Surveill 2020 Oct 22;6(4):e22400 [FREE Full text] [doi: 10.2196/22400] [Medline: 33090117]
5. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018 Nov 01;178(11):1544-1547 [FREE Full text] [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
6. Gindidis S, Stewart S, Roodenburg J. A systematic scoping review of adolescent mental health treatment using mobile apps. Adv Ment Health 2019;17(2):161-177 [FREE Full text] [doi: 10.1080/18387357.2018.1523680]
7. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
8. Singh VK, Long T. Automatic assessment of mental health using phone metadata. Proc Assoc Info Sci Technol 2018;55(1):450-459 [FREE Full text] [doi: 10.1002/pra2.2018.14505501049]
9. Abdullah S, Choudhury T. Sensing technologies for monitoring serious mental illnesses. IEEE MultiMedia 2018 Jan;25(1):61-75 [FREE Full text] [doi: 10.1109/mmul.2018.011921236]
10. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2014 Presented at: UbiComp '14; September 13-17, 2014; Seattle, WA, USA p. 3-14 URL: https://doi.org/10.1145/2632048.2632054 [doi: 10.1145/2632048.2632054]
11. Servia-Rodríguez S, Rachuri KK, Mascolo C, Rentfrow PJ, Lathia N, Sandstrom GM. Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In: Proceedings of the 26th International Conference on World Wide Web. 2017 Presented at: WWW '17; April 3-7, 2017; Perth, Australia p. 103-112 URL: https://doi.org/10.1145/3038912.3052618 [doi: 10.1145/3038912.3052618]
12. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015 Presented at: UbiComp '15; September 7-11, 2015; Osaka, Japan p. 1293-1304 URL: https://doi.org/10.1145/2750858.2805845 [doi: 10.1145/2750858.2805845]

13. Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P. Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 2018;31(4):611-627 [FREE Full text] [doi: 10.1007/s13347-017-0279-x]
14. Kuhlman C, Jackson L, Chunara R. No computation without representation: avoiding data and algorithm biases through diversity. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020 Presented at: KDD '20; July 6-10, 2020; Virtual p. 3593 URL: https://doi.org/10.1145/3394486.3411074 [doi: 10.1145/3394486.3411074]
15. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 2019 Jul 1;63(4/5):4:1-415 [FREE Full text] [doi: 10.1147/JRD.2019.2942287]
16. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD '15; August 10-13, 2015; Sydney, Australia p. 259-268 URL: https://doi.org/10.1145/2783258.2783311 [doi: 10.1145/2783258.2783311]
17. Kamiran F, Karim A, Zhang X. Decision theory for discrimination-aware classification. In: Proceedings of the IEEE 12th International Conference on Data Mining. 2012 Presented at: ICDM '12; December 10-13, 2012; Brussels, Belgium p. 924-929 URL: https://doi.org/10.1109/icdm.2012.45 [doi: 10.1109/icdm.2012.45]
18. Singh VK, Chayko M, Inamdar R, Floegel D. Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. J Assoc Inf Sci Technol 2020 Jan 22;71(11):1281-1294 [FREE Full text] [doi: 10.1002/asi.24335]
19. Bolukbasi T, Chang KW, Zou J, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016 Presented at: NIPS'16; December 5-10, 2016; Barcelona, Spain p. 4356-4364. [doi: 10.5555/3157382.3157584]
20. Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW. Men also like shopping: reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: EMNLP '17; September 7–11, 2017; Copenhagen, Denmark p. 2979-2989 URL: https://doi.org/10.18653/v1/d17-1323 [doi: 10.18653/v1/d17-1323]
21. Rawls J. A Theory of Justice: Revised edition. Cambridge, MA, USA: Harvard University Press; 1999.
22. Lee MK, Jain A, Cha HJ, Ojha S, Kusbit D. Procedural justice in algorithmic fairness: leveraging transparency and outcome control for fair algorithmic mediation. Proc ACM Hum Comput Interact 2019 Nov 07;3(CSCW):1-26 [FREE Full text] [doi: 10.1145/3359284]
23. Duster T. Individual fairness, group preferences, and the California strategy. Representations 1996;55:41-58 [FREE Full text] [doi: 10.2307/3043735]
24. Fish B, Bashardoust A, Boyd D, Friedler S, Scheidegger C, Venkatasubramanian S. Gaps in information access in social networks? In: Proceedings of the 2019 World Wide Web Conference. 2019 Presented at: WWW '19; May 13-17, 2019; San Francisco, CA, USA p. 480-490 URL: https://doi.org/10.1145/3308558.3313680 [doi: 10.1145/3308558.3313680]
25. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. In: Proceedings of the 8th Innovations in Theoretical Computer Science Conference. 2017 Presented at: ITCS '17; January 9-11, 2017; Berkeley, CA, USA URL: https://drops.dagstuhl.de/opus/volltexte/2017/8156/
26. Gummadi KP, Heidari H. Economic theories of distributive justice for fair machine learning. In: Proceedings of the 2019 World Wide Web Conference. 2019 Presented at: WWW '19; May 13-17, 2019; San Francisco, CA, USA p. 1301-1302 URL: https://doi.org/10.1145/3308560.3320101 [doi: 10.1145/3308560.3320101]
27. Ötting SK, Maier GW. The importance of procedural justice in Human–Machine Interactions: intelligent systems as new decision agents in organizations. Comput Human Behav 2018 Dec;89:27-39 [FREE Full text] [doi: 10.1016/j.chb.2018.07.022]
28. Barocas S, Selbst AD. Big data's disparate impact. Calif L Rev 2016;104:671 [FREE Full text] [doi: 10.2139/ssrn.2477899]
29. EEOC v. Sambo's of Georgia, Inc., 530 F. Supp. 86 (N.D. Ga. 1981): US District Court for the Northern District of Georgia. Justia US Law. 1981 Dec 30. URL: https://law.justia.com/cases/federal/district-courts/FSupp/530/86/1370384/ [accessed 2021-07-09]
30. Ware Jr JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992 Jun;30(6):473-483. [Medline: 1593914]
31. Pessach D, Shmueli E. A review on fairness in machine learning. ACM Comput Surv 2023 Apr 30;55(3):1-44 [FREE Full text] [doi: 10.1145/3494672]
32. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16:321-357 [FREE Full text] [doi: 10.1613/jair.953]
33. Monfreda M. Principal component analysis: a powerful interpretative tool at the service of analytical methodology. In: Sanguansat P, editor. Principal Component Analysis. London, UK: IntechOpen; 2012.
34. Dantas J. The importance of k-fold cross-validation for model prediction in machine learning. Towards Data Science. 2020 Nov 4. URL: https://towardsdatascience.com/the-importance-of-k-fold-cross-validation-for-model-prediction-in-machine-learning-4709d3fed2ef [accessed 2021-10-13]

35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine learning in Python. J Mach Learn Res 2011 Oct;12(2011):2825-2830 [FREE Full text]
36. Webster J. Shaping Women's Work: Gender, Employment and Information Technology. Milton Park, UK: Routledge; 2014.
37. Sarwono BK. Gender bias in a patriarchal society: a media analysis on virginity and reproductive health. Wacana 2012 Apr 01;14(1):37-60 [FREE Full text] [doi: 10.17510/wjhi.v14i1.48]
38. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. 2012 Presented at: ITCS '12; January 8-10, 2012; Cambridge, MA, USA p. 214-226 URL: https://doi.org/10.1145/2090236.2090255 [doi: 10.1145/2090236.2090255]
39. Alasadi J, Al Hilli A, Singh VK. Toward fairness in face matching algorithms. In: Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia. 2019 Presented at: FAT/MM '19; October 25, 2019; Nice, France p. 19-25 URL: https://doi.org/10.1145/3347447.3356751 [doi: 10.1145/3347447.3356751]
40. Draws T, Szlávik Z, Timmermans B, Tintarev N, Varshney KR, Hind M. Disparate impact diminishes consumer trust even for advantaged users. In: Proceedings of the 16th International Conference on Persuasive Technology. 2021 Presented at: PERSUASIVE '21; April 12-14, 2021; Virtual p. 135-149 URL: https://doi.org/10.1007/978-3-030-79460-6_11 [doi: 10.1007/978-3-030-79460-6_11]
41. Hogg RV, McKean JW, Craig AT. Introduction to Mathematical Statistics. 6th edition. New York, NY, USA: Pearson; 2005.
42. Kimbrough AM, Guadagno RE, Muscanell NL, Dill J. Gender differences in mediated communication: women connect more than do men. Comput Human Behav 2013 May;29(3):896-900 [FREE Full text] [doi: 10.1016/j.chb.2012.12.005]
43. Forgays DK, Hyman I, Schreiber J. Texting everywhere for everything: gender and age differences in cell phone etiquette and use. Comput Human Behav 2014 Feb;31:314-321 [FREE Full text] [doi: 10.1016/j.chb.2013.10.053]
44. Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 2017 Jun;5(2):153-163. [doi: 10.1089/big.2016.0047] [Medline: 28632438]
45. Taylor WF. The ECOA and disparate impact theory: a historical perspective. J Law Policy 2018;26(2):575 [FREE Full text]
46. Jones J. 5 Facts About the State of the Gender Pay Gap. U.S. Department of Labor Blog. 2021 Mar 19. URL: https://blog.dol.gov/2021/03/19/5-facts-about-the-state-of-the-gender-pay-gap [accessed 2021-10-13]
47. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of 2018 Machine Learning Research Conference on Fairness, Accountability and Transparency. 2018 Presented at: PMLR '18; February 23-24, 2018; New York, NY, USA p. 1-15 URL: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
48. Prates MO, Avelar PH, Lamb LC. Assessing gender bias in machine translation: a case study with Google Translate. Neural Comput Applic 2020;32(10):6363-6381 [FREE Full text] [doi: 10.1007/s00521-019-04144-6]
49. D'Ignazio C, Klein LF. Data Feminism. Cambridge, MA, USA: MIT Press; 2020.
50. Noriega-Campero A, Bakker MA, Garcia-Bulle B, Pentland AS. Active fairness in algorithmic decision making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019 Presented at: AIES '19; January 27-28, 2019; Honolulu, HI, USA p. 77-83 URL: https://doi.org/10.1145/3306618.3314277 [doi: 10.1145/3306618.3314277]
51. Friedman B. Value-sensitive design. interactions 1996 Dec;3(6):16-23 [FREE Full text] [doi: 10.1145/242485.242493]
52. Braveman P, Arkin E, Orleans T, Proctor D, Acker J, Plough A. What is health equity? Behav Sci Policy 2018;4(1):1-14 [FREE Full text] [doi: 10.1353/bsp.2018.0000]
53. Zenou Y, Boccard N. Racial discrimination and redlining in cities. J Urban Econ 2000 Sep;48(2):260-285 [FREE Full text] [doi: 10.1006/juec.1999.2166]
54. Augsberger A, Yeung A, Dougher M, Hahm HC. Factors influencing the underutilization of mental health services among Asian American women with a history of depression and suicide. BMC Health Serv Res 2015 Dec 08;15:542 [FREE Full text] [doi: 10.1186/s12913-015-1191-7] [Medline: 26645481]

Abbreviations

ΔFPR: delta false positive rate
ΔTPR: delta true positive rate
BER: balanced error rate
DIR: disparate impact remover
IMEI: International Mobile Equipment Identity
ML: machine learning
RQ: research question
SAT: Scholastic Assessment Test
SPD: statistical parity difference

Edited by A Mavragani; submitted 19.10.21; peer-reviewed by SM Ayyoubzadeh, E Sükei; comments to author 17.01.22; revised version received 27.03.22; accepted 10.04.22; published 14.06.22

Please cite as:
Park J, Arunachalam R, Silenzio V, Singh VK
Fairness in Mobile Phone–Based Mental Health Assessment Algorithms: Exploratory Study
JMIR Form Res 2022;6(6):e34366
URL: https://formative.jmir.org/2022/6/e34366
doi: 10.2196/34366
PMID: